DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput

Supplementary Notes

Vadim Demichev[1,2], Christoph B. Messner[2], Spyros I. Vernardis[2], Kathryn S. Lilley[1], Markus Ralser[2,3]

1. Department of Biochemistry, The Milner Therapeutics Institute and Cambridge Systems Biology Centre, University of Cambridge, Cambridge, United Kingdom
2. The Francis Crick Institute, Molecular Biology of Metabolism laboratory, London, United Kingdom
3. Department of Biochemistry, Charité Universitätsmedizin Berlin, Berlin, Germany

1. Scoring of putative elution peaks by DIA-NN

Characteristics (73 total) of the putative elution peaks (matched to the respective target or decoy precursor ions) scored by DIA-NN and used by the neural networks ensemble classifier are summarised in the following table:

Ions co-elution (MS2 level)	
Pearson correlations of top 12 fragments' elution profiles (fragments ordered by their reference library intensities) with the smoothed elution profile of the "best" fragment	12 scores
Sum of these correlations for the top 6 fragments; calculated for chromatograms extracted at the base mass accuracy as well as 0.45 and 0.2 of the base mass accuracy	3 scores
Sum of these correlations for the rest of the fragments, without normalisation and normalised by the number of these extra fragments	2 scores
Sum of these correlations for the 3 b-series charge 1 fragments of the precursor with	1 score; library-free search only

highest correlations (among all such b-series fragments)	
Sum of such correlations for the top 6 fragments with elution profiles x() first processed using the x(scan) -> min(x(scan - 1), x(scan), x(scan + 1)) replacement	1 score; this score is included, as it might be beneficial when using overlapping window workflows, i.e. if isolation windows are shifted, e.g. by half window width, in each subsequent cycle
Correlation between the elution profile at the m/z value that corresponds to the non-fragmented precursor and the smoothed elution profile of the “best” fragment	1 score
Ions co-elution (MS1 level)	
Pearson correlations of the smoothed elution profile of the “best” fragment with the MS1 elution profiles extracted using the base mass accuracy as well as 0.45 and 0.2 of the base mass accuracy	3 scores
Isotopologue ions co-elution	**Use of these scores can be turned off by the user, e.g. if a C13-depleted sample is being analysed**
Pearson correlations of the smoothed elution profile of the “best” fragment with the MS1 elution profiles corresponding to the peptide featuring 1, 2 or 3 C13	3 scores
Sum of Pearson correlations of elution profiles corresponding to the top 6 fragments featuring one C13 with the smoothed elution profile of the “best” fragment	1 score
Pearson correlations of elution profiles corresponding to the top 6 fragments, from masses of which (C13 - C12) mass was subtracted, with the smoothed elution profile of the “best” fragment	6 scores; these scores would reflect it if the fragments observed were actually heavy isotopologues of some fragments with lower monoisotopic masses, thus being unlikely to belong to the peptide
Sum of these correlations	1 score

Total signal	
Natural logarithm of the sum of the areas below the elution curves of the top 6 fragments multiplied by the respective correlations with the smoothed elution profile of the “best” fragment	1 score
Measured relative fragment intensities	
Cosine similarity measure (itself and to power 3) between the predicted and measured intensities of the top 6 fragments weighted by the squared values of the smoothed “best” fragment elution curve at the respective time points	2 scores
Relative intensities of the top 6 fragments	6 scores
Mass accuracy (MS2)	
Measured mass accuracy of the top 6 fragments at the elution apex weighted by the Pearson correlations of the respective elution curves with the smoothed elution curve of the “best” fragment	6 scores
Retention time (RT)	
Retention time apex	1 score
Square root of the absolute difference between measured and predicted retention times	1 score
Elution profile shape	
The chromatogram scanning window is split into five segments and relative total intensities are calculated for the “best” fragment for these segments	5 scores
Presence of other putative elution peaks	

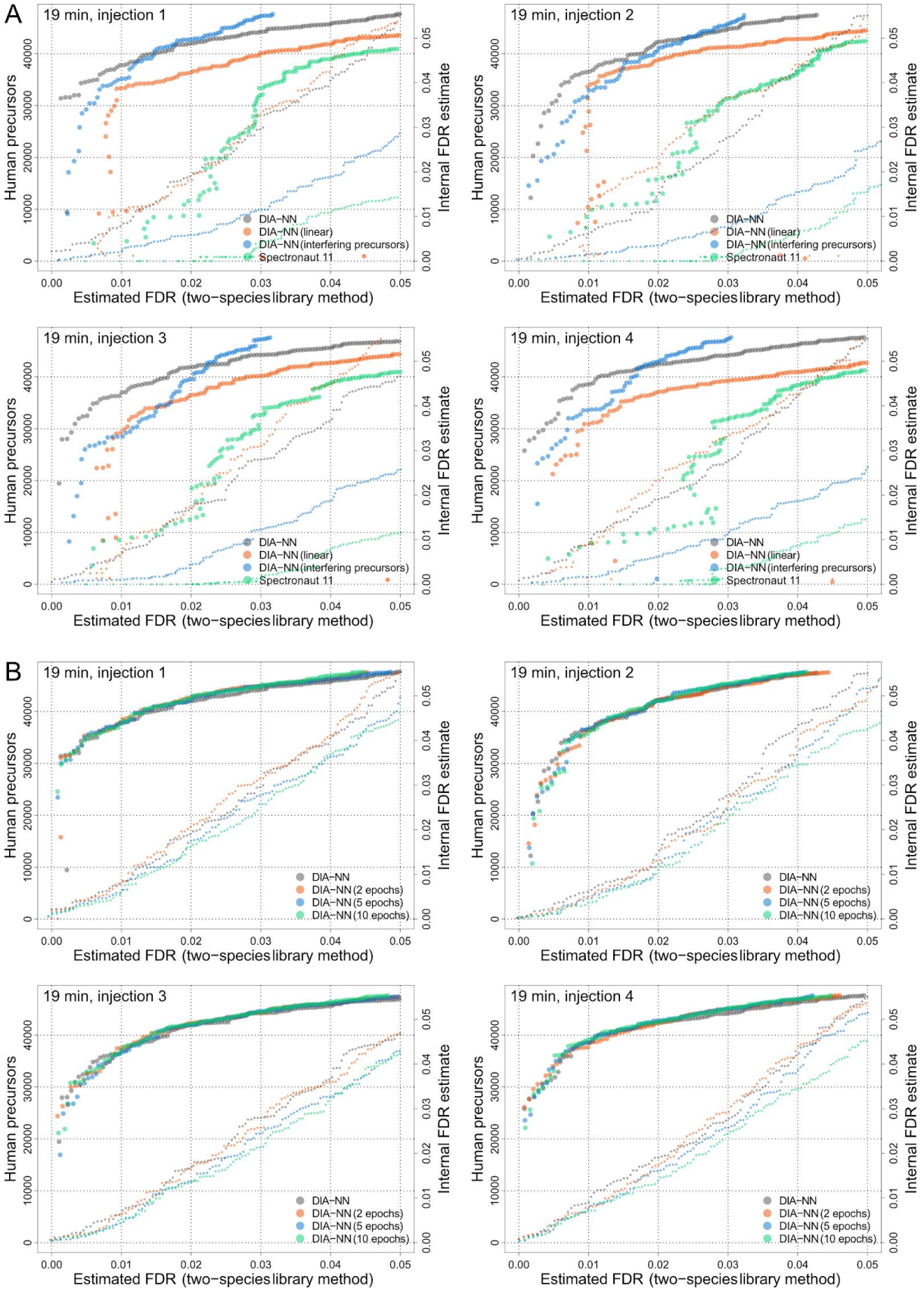
Sum of the Pearson correlations between the elution profiles of the top 6 fragments and the smoothed elution profile of the "best" fragment, from which the maximum of such correlation sums for all putative elution peaks considered has been subtracted	1 score
$log(max(1.0, s) / (S + 1.0))$, where s is the sum of the Pearson correlation between the elution profiles of the top 6 fragments and the smoothed elution profile of the "best" fragment, and S is the sum of such correlation sums for all putative elution peaks	1 score
Library characteristics of the precursor	
Library intensities of the fragments 2 to 12 (fragments ordered by their reference library intensities) relative to the top fragment	11 scores
precursor m/z	1 score
precursor charge	1 score
precursor length	1 score
number of library fragments	1 score

2. The use of non-project specific spectral libraries with DIA-NN

DIA data are increasingly analysed with publicly available spectral libraries, not necessarily generated on the same LC-MS setup. This happens when the creation of a project-specific spectral library of comparable depth is not possible due to low amounts of sample available, or the efforts to create a specific library are not justified in view of the aims of a particular experiment. At the same time, a publicly available library can sometimes be expected to yield better performance than a library-free analysis, e.g. when the samples have been analysed with a very short chromatographic gradient. To demonstrate the performance of DIA-NN in this situation, we used the same human-maize spectral library, as utilised to obtain Figure 1B,

to analyse four consecutive injections of human myelogenous leukemia cell line K562 proteomic preparation (online Methods) analysed on a different LC-MS setup (microflow HPLC coupled to TripleTOF 6600 (Sciex)). A fast (19 minute) chromatographic gradient was used. These acquisitions were processed with Spectronaut Pulsar and DIA-NN using the default settings as well as with the neural network classifier and the removal of potentially interfering precursors disabled (Supplementary Fig. SN2.1A). Protein inference and FDR filtering were turned off, to obtain complete reports, "Unrelated runs" option was checked in the DIA-NN settings. A two-species human-maize spectral library method was used to estimate the effective FDR, as when generating Figure 1B (online Methods).

This example illustrates that non-project specific libraries are effectively used by DIA-NN, even if they have been recorded on a different mass spectrometer and LC setup. Moreover, the example illustrates that DIA-NN consistently retains ID performance; the tool was able to identify several times more precursors at strict FDR compared to Spectronaut. The use of neural networks classifier enables DIA-NN to achieve effective FDR as low as 0.2% in this test. The algorithm that removes interfering precursors, on the other hand, not only improves the identification performance at strict FDR thresholds, but also substantially affects the internal FDR estimates of DIA-NN, bringing them in close agreement with FDR estimates obtained using the two-species library method. We note that in this benchmark DIA-NN identifies more precursors at 1% FDR from a 19 minute microflow gradient than what was achieved recently with 120 minute nanoflow gradient[1]. This highlights the promise fast gradients now hold for high-throughput proteomics, with the development of new fast instruments and advanced software, such as DIA-NN. We also evaluated the degree to which the internal FDR estimates of DIA-NN might be affected by potential overfitting by neural networks, and concluded that at least in this benchmark the effect is either nonexistent or negligible (Supplementary Fig. SN2.1B).

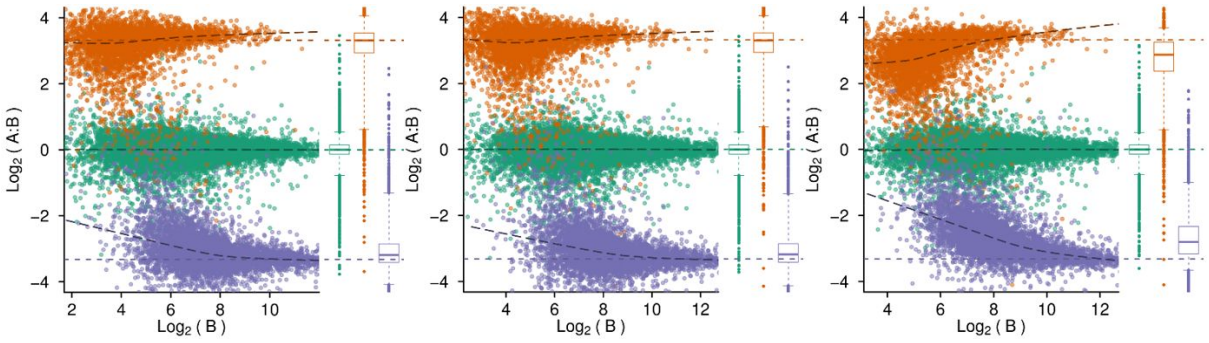


Supplementary Fig. SN2.1. **(A) Efficient peptide identification with non-project specific spectral libraries in high-throughput proteomics**. DIA-NN with default settings, with neural networks classifier disabled ("DIA-NN (linear)") and with removal of interfering precursors disabled ("DIA-NN (interfering precursors)") benchmarked against Spectronaut Pulsar using four consecutive injections of human myelogenous leukemia cell line K562

proteome preparations analysed with a 19 minute microflow chromatographic gradient recorded on a TripleTOF 6600 (Sciex) mass spectrometer. Processing was performed with a (non-project-specific) two-species human-maize spectral library (online Methods), generated on a Q Exactive HF coupled with a nanoLC (the same library was used to produce Figure 1B). Identification numbers (large points, left y-axis) and internal FDR estimates (small points, right y-axis) are plotted against the estimates of the effective FDR using the two-species method (online Methods). DIA-NN substantially outperforms Spectronaut at strict FDR (<2%), identifying several times more precursors, and demonstrates more accurate internal FDR estimates. **(B) Single training epoch (default) is unlikely to lead to overfitting**. We benchmarked the performance of DIA-NN when training the neural network for a single epoch (default) or up to 10 epochs. We observe that some overfitting starts being noticeable only at about 10 epochs (manifested as slightly optimistic FDR estimates), but even 10 epochs are not enough to cause a substantial change in the accuracy of FDR estimates. We also observe that the performance measured with the use of a two-species human-maize spectral library is not affected by the number of training epochs, indicating that the choice of only a single epoch as the default setting is sensible, at least for this kind of data.

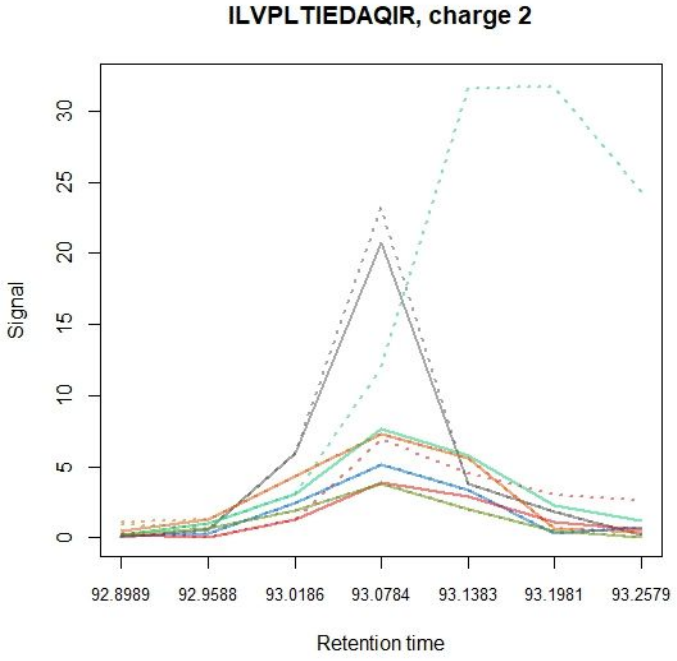
3. Performance benchmark of DIA-NN's quantification algorithms

In this section we use LFQbench to evaluate the effectiveness of DIA-NN's quantification algorithms (Supplementary Fig. SN3.1). We see that while cross-run selection of fragments for quantification does not have a noticeable impact on quantification accuracy in this test, the removal of interferences from fragment elution profiles significantly improves quantification of low-abundant peptides, resulting in more accurate quantification ratios between different species lysate mixtures (A and B). We also provide an illustration of the way interference removal works (Supplementary Fig. SN3.2; see Methods for the algorithmic details).



Supplementary Fig. SN3.1. **Performance of DIA-NN's quantification algorithms.** LFQbench peptide ratio plots for DIA-NN in the default configuration (left), with the cross-run selection of fragments for quantification disabled (middle) and with removal of interferences from fragments' elution profiles also disabled (right). In box and whisker plots,

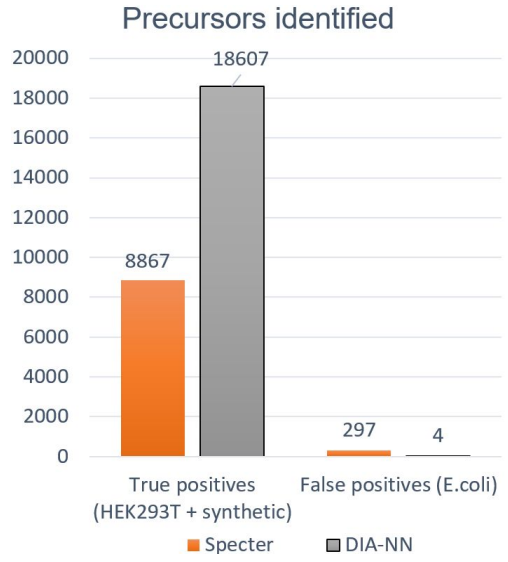
boxes correspond to interquartile ranges and whiskers to 1-99 percentiles; n = 15743 (human), 3755 (yeast), 4997 (*E.coli*).



Supplementary Fig. SN3.2. **Removal of interferences from fragments' elution curves by DIA-NN.** Chromatograms (LFQbench test, acquisition 1A) for an *E.coli* peptide are plotted before (dotted lines) and after (solid lines) interference correction. DIA-NN assumed that the third library fragment (m/z 487.299, blue elution curve) was representative of the true elution curve of the peptide, and used its extracted elution profile to remove interferences from the extracted elution profiles of other fragments.

4. Benchmark of DIA-NN against Specter[2].

We also compared the identification performance of DIA-NN with that of Specter, using the benchmark method from the original manuscript[2]. To do this, we employed the HEK293T dataset (with spiked-in synthetic peptides; PXD006722 ProteomeXChange repository) and the two-species spectral library (human - *E.coli*) described previously[2] (Supplementary Fig. SN4.1). Of note, we used the identification numbers reported for Specter therein (Figure 3a[2]), meaning that the parameters of Specter were optimised by the Specter authors themselves.

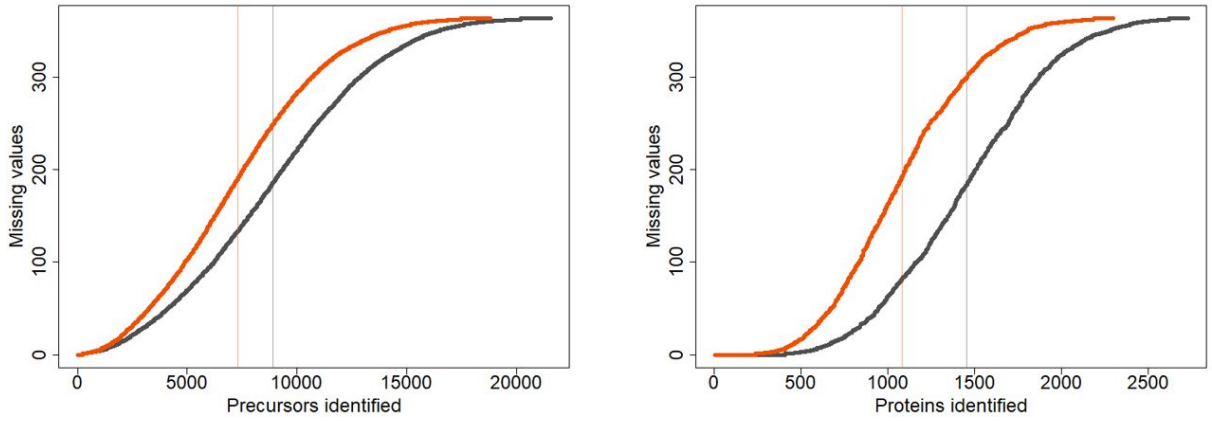


Supplementary Fig. SN4.1. **Comparison of the identification performance of DIA-NN and Specter**. DIA-NN was used to reanalyse the three HEK293T injections (with spiked-in synthetic peptides; "6.75ng", acquired using a 53 min chromatographic gradient on Q Exactive HF Plus) described previously[2] with the respective project-specific spectral library concatenated with an *E.coli* spectral library[2]. The libraries were converted from the .blib (Skyline) format to a simple text format using the Specter code; DIA-NN was then used to annotate the fragments (with mass accuracy set to 5 ppm, maximum fragment charge set to 2, and H_2O as well as NH_3 neutral losses allowed) and exclude precursors with less than 6 fragments annotated (generating a library of 28633 HEK293T + synthetic precursors and 47261 *E.coli* precursors). Analysis using the resulting library was performed with the default DIA-NN settings, except the precursor q-value threshold was set to 0.01% (i.e. 0.0001). The numbers of cumulatively identified in three technical replicates precursors were then calculated. The respective numbers for Specter were taken out of the Specter original manuscript[2]. DIA-NN identified more than 2x greater number of precursors with estimated effective FDR (proportional to the ratio of the number of reported *E.coli* identifications and the number of all identifications) being over two orders of magnitude (155x) lower.

5. Consistency of the neural networks classifier

DIA-MS proteomics is known for its high consistency of identification, leading to fewer missing values in comparison to DDA-based approaches. Here we show that the use of a deep neural networks (DNNs) classifier does not have a negative impact on the identification consistency when analysing multiple acquisitions. To demonstrate this, we used 364 yeast proteomes generated by us previously[3] (PXD010529 ProteomeXChange repository) and analysed this dataset using a spectral library generated from DIA data (as described below, see Supplementary Note 10) at 1% q-value (as reported by DIA-NN), with the DNNs enabled and disabled (Supplementary Fig. SN5.1).

We note that these data were acquired on a previous generation instrument (Sciex TripleTOF 5600) with a workflow optimised to provide high precision of quantification for highly-abundant enzymes, without regard for the identification performance[4]. The identification numbers are thus lower than what is possible now, e.g. with a modern workflow used to generate Supplementary Fig. SN10.1.



Supplementary Fig. SN5.1. **The use of the deep neural networks does not have a negative impact on identification consistency.** The number of missing values (out of 364 acquisitions) plotted against the numbers of precursors (left) and proteins (right; only uniquely identified proteins considered) with deep neural networks classifier enabled (black) and disabled (orange). The respective mean identification numbers are indicated with vertical lines.

6. Hardware requirements, speed and GUI

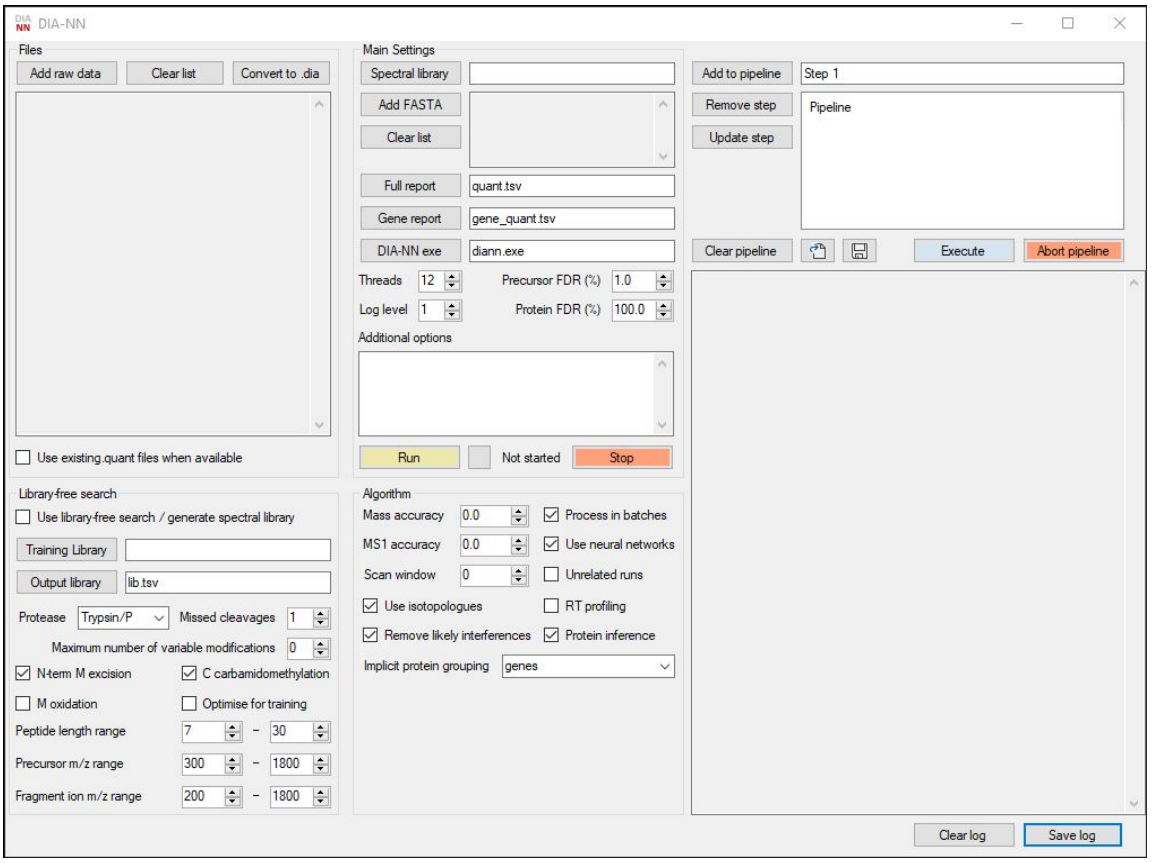
The rising interest in high-throughput proteomics in research, medicine and industry calls for the development of software tools that are able to rapidly and reliably analyse thousands of mass spectrometry acquisitions. DIA-NN performs the computationally-demanding processing steps separately for each acquisition in the experiment, saving all the relevant information to compact files on the hard drive. This allows quick and flexible analysis and subsequent reanalysis of any part of the experiment separately. DIA-NN is also very fast (Supplementary Table SN6.1).

Figure	Figure 1B	Supplementary Figure SN2.1	Supplementary Figure SN8.2
Dataset	4 HeLa acquisitions	4 K562 acquisitions	6 mixed-species acquisitions
Analysis mode	spectral library	spectral library	library-free
Spectronaut: input file format	.raw	.htrms	.wiff
DIA-NN: input file format	.raw	.dia	.dia
Spectronaut: total time	26 min	42 min	22 hours
Spectronaut: RAM, peak working set	15 GB	23 GB	13 GB
DIA-NN: total time	10 min	7.5 min	10 hours
DIA-NN: RAM, peak working set	3.3 GB	1.6 GB	11 GB

Figure	Supplementary Figure SN10.1	Supplementary Figure SN5.1
Dataset	3 yeast acquisitions	364 yeast acquisitions
Analysis mode	library-free	spectral library
Input file format	.dia	.dia
DIA-NN: total time	26 min	3.3 hours
DIA-NN: RAM, peak working set	2.4 GB	1.6 GB

Supplementary Table SN6.1. **Processing speed of DIA-NN.** Time and computer memory amount required by DIA-NN (both panels) and Spectronaut (upper panel) to process the datasets featured in the respective Figures from the present manuscript. The data for Spectronaut (unlike for DIA-NN) does not include the time required for report generation. Benchmarks were conducted using a PC based on 2x Xeon X5650 (2x 6-core, 2.67GHz) under Windows 7, RAM usage was measured using the Windows Task Manager, results were rounded to two significant figures.

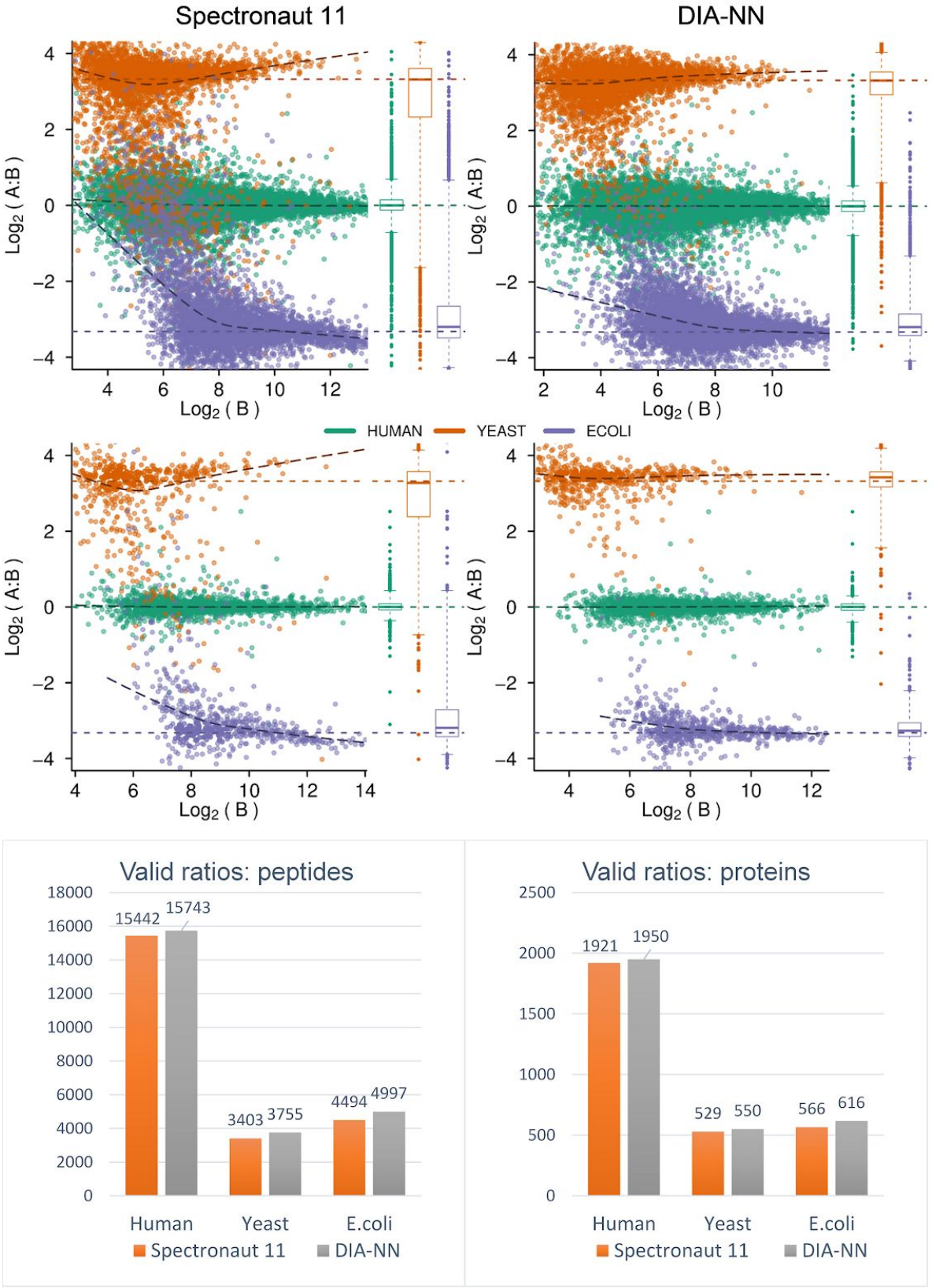
For large scale applications, we provide a command line tool, which can be used to set up automatic processing workflows. For smaller or more routine applications, we have further programmed a graphical user interface (GUI) wrapper, that enables the control of all steps of the workflow from a simple and intuitive workspace (Supplementary Fig. SN6.2). Although DIA-NN is designed to do as much as possible automatically, it is fully configurable, allowing to fine-tune the processing workflow for a specific experiment. The GUI is implemented as a wrapper for the command line tool, and thus allows to easily set up the analysis in few clicks without losing the powerful tuning capabilities of the command line tool: the GUI is capable of carrying out any task supported by the command-line tool, provided it involves analysis (of an arbitrary number of experiments) using parameters and input files defined by the user. The command line tool, however, can also be invoked (by custom scripts) using dynamically-generated data, e.g. it can be used to convert or process “on the fly” the raw data files being acquired by the LC-MS.



Supplementary Fig. SN6.2. **DIA-NN graphical user interface**. DIA-NN allows to control the full workflow from a simple to use and intuitive graphical user interface (GUI). The GUI supports fully automatic processing, but also allows to fine-tune the algorithm. Pipelines can be constructed to automatically process multiple experiments with different settings.

7. Performance of DIA-NN in the LFQbench test

While the identification performance is important, so far the key application of DIA is accurate, precise and consistent peptide and protein quantification in large sample series. We illustrated the quantification performance of DIA-NN by comparing it to Spectronaut Pulsar using the LFQbench test[5] (HYE110 dataset, 64 variable window acquisition scheme on TripleTOF 6600) (Supplementary Fig. SN7.1). In this benchmark, human, yeast, and *E.coli* lysates were mixed in different proportions and analysed via SWATH-MS. For each mixture, three injection replicates were measured. The performance of the software tools was compared using the LFQbench R package (https://github.com/IFIproteomics/LFQbench), which takes as input the intensities of the precursor ions and uses these to quantify peptides and proteins. Q-value threshold was set to 1% (note that actual FDR might differ substantially for DIA-NN and Spectronaut in this test, as these tools have different FDR estimation algorithms). The default settings were used for DIA-NN and Spectronaut, except that protein inference and FDR filtering of the output were turned off to obtain complete reports.



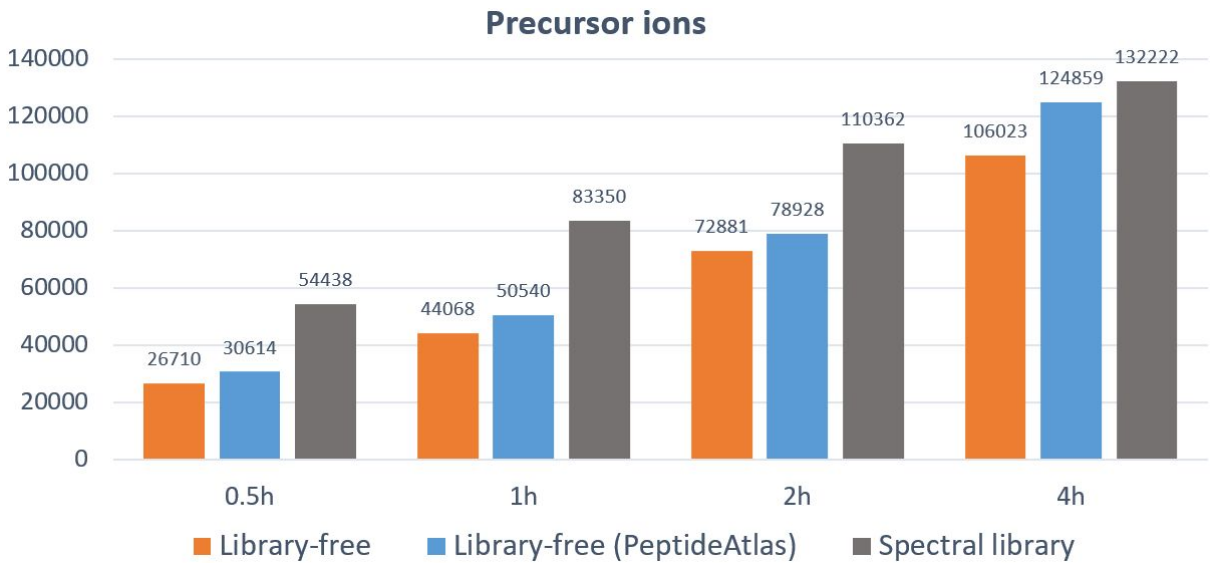
Supplementary Fig. SN7.1. **Performance of DIA-NN in the LFQbench test (Complete Figure, of which an extract is shown in Figure 2)**. LFQbench performance of DIA-NN in comparison to Spectronaut. In the LFQbench test, two peptide preparations (yeast and *E.coli*) are mixed in two different proportions (A and B), pooled with a human peptide preparation and analysed in triplicates on TripleTOF 6600[5] (64-variable windows acquisition, HYE110 dataset). The data were processed at 1% precursor q-value; peptide (top panel) and protein

(middle panel) ratios between the mixtures were visualised using the LFQbench R package (with the dashed lines indicating the expected ratios). DIA-NN demonstrates significantly better quantification precision for both yeast and *E.coli* peptides and proteins, as evidenced by the box plots for the ratios (boxes: interquartile range, whiskers: 1-99 percentile; n = 15442 and 15743 (human), 3403 and 3755 (yeast), 4494 and 4997 (*E.coli*) for peptide ratios obtained from the reports of Spectronaut and DIA-NN, respectively; n = 1921 and 1950 (human), 529 and 550 (yeast), 566 and 616 (*E.coli*) for protein ratios obtained from the reports of Spectronaut and DIA-NN, respectively). DIA-NN also produced better median CV values for human peptides and proteins: 5.6% and 3.0%, respectively, compared to 7.0% and 3.8% for Spectronaut, as calculated by the LFQbench R package. Bottom panel: numbers of valid A:B ratios produced on the peptide and protein level.

8. Library-free processing

DIA-NN can process raw data using either a spectral library or a protein sequence database. In the latter case, proteins are *in silico* digested and prediction of the fragmentation spectra of the resulting peptides as well as the respective retention times is performed (Methods). Further processing is done by the same algorithms as in the spectral library-based search, meaning that DIA-NN's library-free module is largely a peptide-centric search tool, similarly to PECAN[6], in contrast to e.g. DIA-Umpire[7], which utilises a spectrum-centric approach.

While the spectral library-based search achieves higher proteomic depth, the library-free approach saves sample material and the instrument time. We benchmarked the library-free performance of DIA-NN using HeLa proteome analyses obtained with different chromatographic gradient lengths[8] (Supplementary Fig. SN8.1). Library-free processing was carried out against the human UniProt[9] canonical proteome (3AUP000005640). To demonstrate the benefit of restricting the search space, the data were also processed with the same sequence database but filtered to include only peptides known to be present in human samples, according to the PeptideAtlas[10] build of January 2018; for this, the maximum peptide length was set to 100 and up to five missed cleavages were allowed. An *E.coli* spectral library[8] was used to train the peptide fragmentation and retention time predictors. A project-specific spectral library[8] was used for library-based processing.



Supplementary Fig. SN8.1. **Library-free performance of DIA-NN**. The numbers of precursors identified at 1% q-value threshold, as reported by DIA-NN for HeLa acquisitions on Q Exactive HF as a function of chromatographic gradient length.

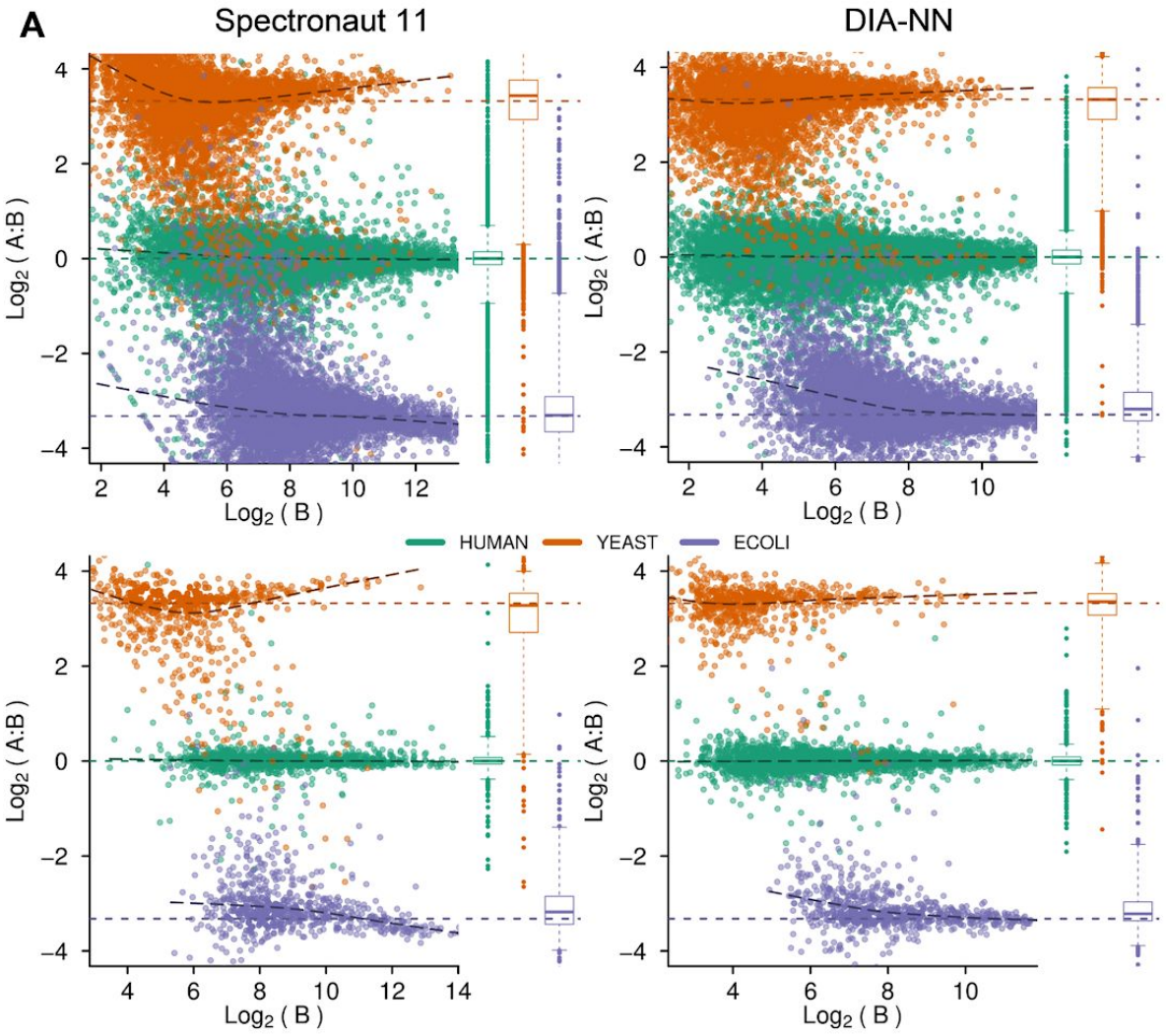
To validate the library-free performance of DIA-NN, we used the LFQbench[5] dataset, as when generating Figure 2 and Supplementary Fig. SN7.1. However, instead of using the DDA-based spectral library provided, we employed a library-free workflow. In general, we would recommend creating a spectral library directly from DIA data acquired on the same LC-MS setup using gas-phase fractionation. However, as such a library is not available for the LFQbench dataset, we utilised a two-step procedure analogous to the approach suggested for DIA-Umpire[7]. To do this automatically, we employed the pipeline capability of the DIA-NN GUI. First, DIA-NN was used to produce a spectral library from the 6 acquisitions included in the LFQbench dataset, by searching against the human, yeast and *E.coli* UniProt[9] canonical proteomes (with the respective IDs 3AUP000005640, 3AUP000002311 and 3AUP000000625) at the same time (the LFQbench spectral library was used to train the peptide fragmentation and retention time predictors). The same acquisitions were then reanalysed with this library. The q-value threshold was set to 0.5% at the precursor level for both of these steps. Quantification performance was assessed using the LFQbench R package. For comparison, we also performed directDIA analysis of the same dataset with Spectronaut with its default settings. Default q-value filtering (1% precursor level and 1% protein level) was used for Spectronaut, as we found that relaxing it in library-free mode has a detrimental effect on the performance of Spectronaut. We note that similarly to the two-step DIA-NN workflow, Spectronaut also enhances identification performance by processing multiple

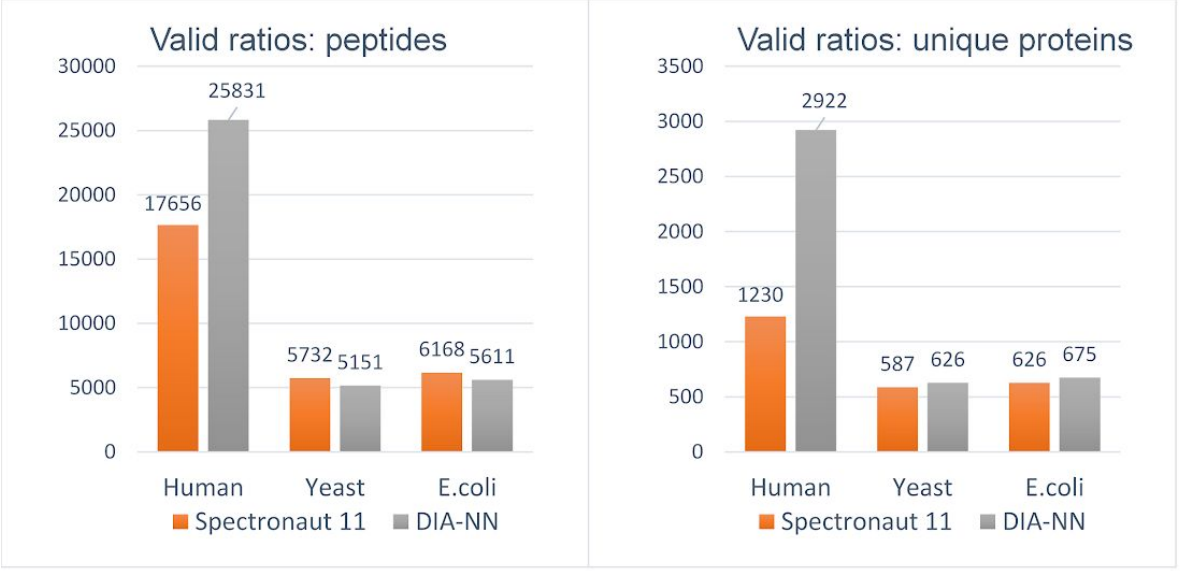
acquisitions at once, reporting more identifications than when processing the same acquisitions separately. LFQbench R package was directly used to analyse the reports of DIA-NN and Spectronaut, with the respective protein names columns used to infer the species for each precursor.

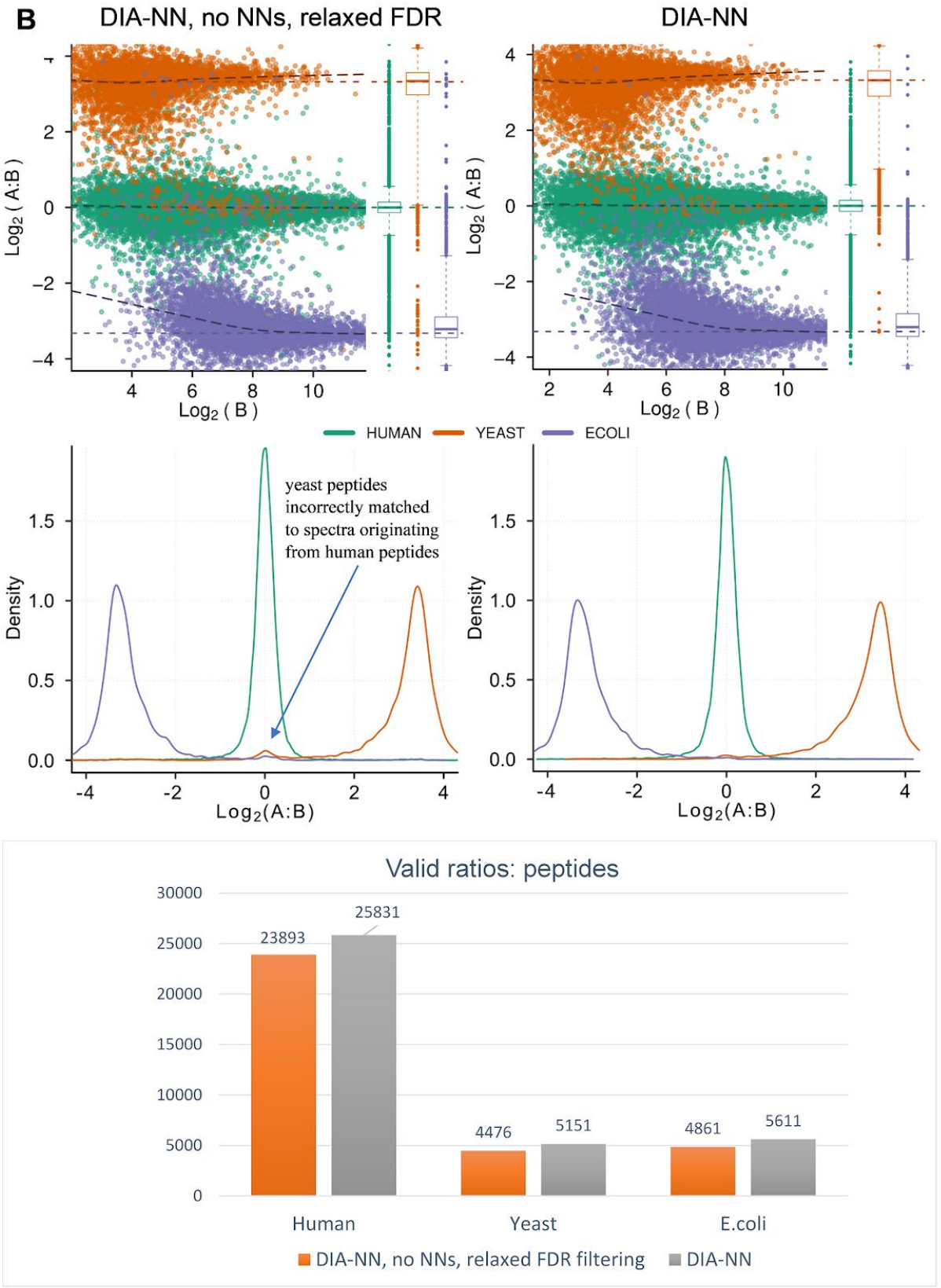
Our results (Supplementary Fig. SN8.2A) demonstrate that DIA-NN identifies substantially more peptides in this library-free workflow than when using the spectral library generated by Navarro et al[5] (cf. Supplementary Fig. SN7.1). DIA-NN also identifies substantially more human peptides than Spectronaut, while the median CVs were reported to be 7.2% for both DIA-NN and Spectronaut, i.e. DIA-NN is able to quantify substantially more peptides at the same precision. These results are thus in line with the superior quantification capabilities of DIA-NN demonstrated when using a DDA-based spectral library (Figure 2 and Supplementary Fig. SN7.1).

We also plotted the A:B ratios for proteins uniquely identified by DIA-NN and Spectronaut (Supplementary Fig. SN8.2A, middle panel). For this, the reports produced by DIA-NN and Spectronaut were filtered to include only precursors matched (by their respective protein grouping algorithms) to a single protein. Such filtering is often performed in practice, with precursors associated with multiple proteins being discarded. In this test, DIA-NN quantified more human, yeast and *E.coli* proteins, while demonstrating better quantification performance (Supplementary Fig. SN8.2A). Of note, the numbers of valid protein ratios obtained in this test should not be directly compared to those obtained previously: when analysing with a DDA-based library (Supplementary Fig. SN7.1) it was protein groups, rather than unique proteins, that were counted.

Finally, we validate the advantage of neural networks over the linear classifier in library-free mode (Supplementary Fig. SN8.2B).


A
Spectronaut 11
DIA-NN
HUMAN
YEAST
ECOLI
Log2 (A:B)
Log2 (B)


Valid ratios: peptides
30000
25000
20000
15000
10000
5000
0
25831
17656
5732
5151
6168
5611
Human
Yeast
E.coli
Spectronaut 11
DIA-NN
Valid ratios: unique proteins
3500
3000
2500
2000
1500
1000
500
0
2922
1230
587
626
626
675
Human
Yeast
E.coli
Spectronaut 11
DIA-NN



Supplementary Fig. SN8.2. **(A) Library-free performance of DIA-NN in the LFQbench test.** Quantification ratios between A and B species mixtures for library-free analysis of the LFQbench dataset by Spectronaut (left) and DIA-NN (right) plotted at the levels of peptides (top panel) and uniquely identified proteins (middle panel). Library-free search was performed against the UniProt[9] canonical proteomes 3AUP000005640 (human), 3AUP000002311 (yeast) and 3AUP000000625 (*E.coli*). Peptide ratios between the mixtures

were visualised using the LFQbench R package (boxes: interquartile range, whiskers: 1-99 percentile; n = 17656 and 25831 (human), 5732 and 5151 (yeast), 6168 and 5611 (*E. coli*) for peptide ratios obtained from the reports of Spectronaut and DIA-NN, respectively; n = 1230 and 2922 (human), 587 and 626 (yeast), 626 and 675 (*E.coli)* for protein ratios obtained from the reports of Spectronaut and DIA-NN, respectively). Bottom panel: the respective numbers of valid ratios. **(B) Neural networks improve DIA-NN's performance in library-free mode.** LFQbench dataset was analysed by DIA-NN in library-free mode with neural networks disabled (2% q-value filtering) and enabled (0.5% q-value filtering, as in **A**). The q-value threshold of 2% was chosen to keep the resulting library size less than when using neural networks and filtering at 0.5% (50032 vs 52912 precursors in the generated library, respectively). This way we expected the advantage of neural networks to manifest as simultaneously slightly higher number of valid A:B ratios obtained for the peptides as well as several times lower number of grossly incorrect ratios, which are expected to be indicative of false identifications. This is indeed the case. The top panel corresponds to the peptide ratio plots and middle panel to ratio distributions, as generated by LFQbench R package for DIA-NN without neural networks (left) and with neural networks (right). Peptide ratios between the mixtures were visualised using the LFQbench R package (boxes: interquartile range, whiskers: 1-99 percentile; n = 23893 and 25831 (human), 4476 and 5151 (yeast), 4861 and 5611 (E.coli), without and with neural networks, respectively). The bottom panel presents the numbers of valid ratios.

9. Precursor ions removed from the human-maize spectral library

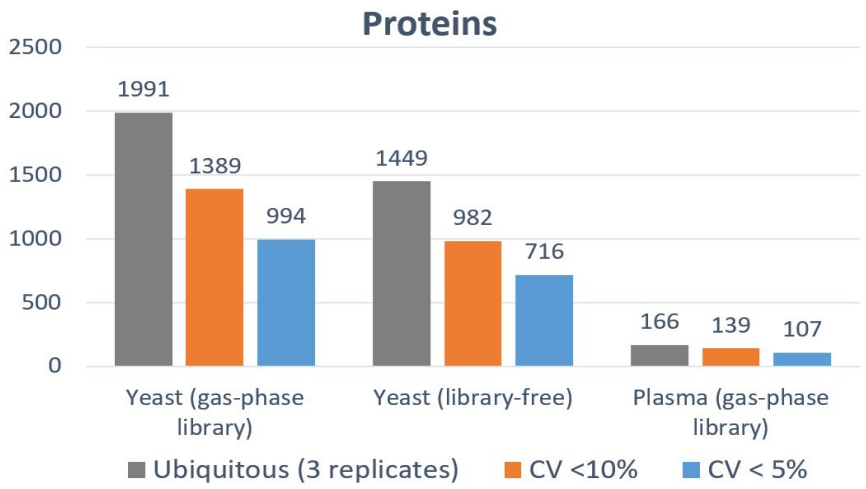
The following peptides were removed from the two-species human-maize spectral library, in order to facilitate its import into Skyline. Peptide modifications are encoded in the UniMod format, precursor charges are indicated with a number following the amino acid sequence.

UniMod:1)M(UniMod:35)M(UniMod:35)GHRPVLVLSQN(UniMod:7)TK3
(UniMod:1)SADGAEADGSTQVTVEEPVQQ(UniMod:7)PSVVDR3
(UniMod:1)SAPLDAALHALQEEQ(UniMod:7)AR2
(UniMod:1)SELDQLRQEAEQ(UniMod:7)LK2
(UniMod:1)SELEQ(UniMod:7)LRQEAEQ(UniMod:7)LR2
(UniMod:1)SELEQLRQEAEQ(UniMod:7)LR2
(UniMod:1)SGEENPASKPTPVQDVQ(UniMod:7)GDGR2
(UniMod:1)SHVAVENALGLDQ(UniMod:7)QFAGLDLNSSDNQSGGSTASK3
(UniMod:1)SHVAVENALGLDQQ(UniMod:7)FAGLDLNSSDNQSGGSTASK3
(UniMod:1)SHVAVENALGLDQQFAGLDLN(UniMod:7)SSDNQSGGSTASK3
(UniMod:1)SHVAVENALGLDQQFAGLDLNSSDN(UniMod:7)QSGGSTASK3
(UniMod:1)SHVAVENALGLDQQFAGLDLNSSDNQ(UniMod:7)SGGSTASK3
(UniMod:1)SKPHSEAGTAFIQTQQ(UniMod:7)LHAAMADTFLEHM(UniMod:35)C(UniMod:4)R5
(UniMod:1)SLIC(UniMod:4)SISNEVPEHPC(UniMod:4)VSPVSN(UniMod:7)HVYER3
(UniMod:1)SQ(UniMod:7)DGASQFQ(UniMod:7)EVIR2
(UniMod:1)SQDGASQFQ(UniMod:7)EVIR2
(UniMod:1)STGTFVVSQPLN(UniMod:7)YR2
(UniMod:1)STLLINQPQ(UniMod:7)YAWLK2
(UniMod:1)STNEN(UniMod:7)ANTPAAR2

(UniMod:1)STNENAN(UniMod:7)TPAAR2
(UniMod:1)STSVPQGHTWTQ(UniMod:7)R2
(UniMod:1)TSALENYIN(UniMod:7)R2
(UniMod:1)TTQQ(UniMod:7)IDLQGPGPWGFR2
(UniMod:1)TTYLEFIQQ(UniMod:7)NEER2

10. Generating spectral libraries with DIA-NN

DIA-NN can generate spectral libraries directly from DIA data. Here, we demonstrate its capabilities using a workflow optimised for high-throughput proteome quantification based on 21 to 23-minute gradient microflow SWATH[11] applied on yeast and human plasma proteomes (Supplementary Fig. SN10.1). Briefly, Sciex TripleTOF 6600 was used to rapidly analyse yeast and human plasma tryptic digests (three injections each; see the detailed workflow description in the online Methods section). Furthermore, a set of SWATH gas-phase fractionation acquisitions with narrow precursor isolation windows was acquired for each of the digests. DIA-NN was used to create DIA-based spectral libraries directly from these gas-phase fractionation acquisitions. For the yeast library, a search against the yeast UniProt[9] canonical proteome was used (3AUP000002311). For the plasma library, the acquisitions were searched against the human UniProt canonical proteome (3AUP000005640) filtered for the peptides known to be present in human plasma, according to the PeptideAtlas[10] build of August 2013; for this, the maximum peptide length was set to 100 and up to five missed cleavages were allowed. The numbers of proteins uniquely identified in all three technical replicates at 1% q-value (i.e. using peptides specific to the respective genes) were then calculated, as well as the numbers of these with coefficients of variation (CV) less than the specified thresholds. Yeast acquisitions were also analysed by DIA-NN directly, without the DIA-based spectral library. For the initial analysis of the yeast gas-phase fractionation acquisitions, which was performed at 0.2% precursor q-value, an *E.coli* spectral library[8] was used to train the peptide fragmentation and retention time predictors, with the "Optimise for training" option selected in the DIA-NN GUI. Subsequently, all library-free processing was performed using the newly generated library, specific to our LC-MS setup, to train the predictors. The 1% precursor q-value threshold was used for the rest of the analyses.



Supplementary Fig. SN10.1. **Using DIA-NN to analyse yeast and human plasma SWATH acquisitions without a DDA-based spectral library**. DIA-NN was used to analyse yeast (23-minute gradient) and human plasma (21-minute gradient) triplicate acquisitions using spectral libraries generated by DIA-NN from gas-phase fractionation acquisitions. A library-free analysis of the same yeast acquisitions was added for comparison. Only uniquely identified proteins (i.e. using proteotypic peptides only), detected in all three replicates, were considered (filtered at 1% precursor-level and 1% protein-level q-value).

References

1. Bruderer, R. *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410 (05/2015).
2. Peckner, R. *et al.* Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nat. Methods* **15**, 371–378 (2018).
3. Zelezniak, A. *et al.* Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts. *Cell Syst* **7**, 269–283.e6 (2018).
4. Vowinckel, J. *et al.* The beauty of being (label)-free: sample preparation methods for SWATH-MS and next-generation targeted proteomics. *F1000Res.* (2014). doi:10.12688/f1000research.2-272.v2
5. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).
6. Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**, 903–908 (2017).
7. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
8. Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. Cell. Proteomics* **16**, 2296–2309 (12/2017).
9. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
10. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9**, 429–434 (2008).
11. Vowinckel, J. *et al.* Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. *Sci. Rep.* **8**, 4346 (2018).